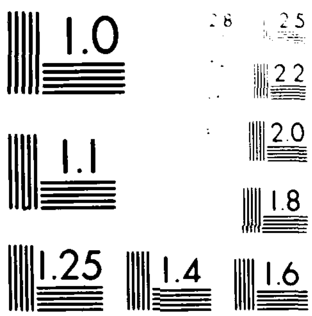
AD-A116 190 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
RECURSIVE PARAMETER ESTIMATION USING INCOMPLETE DATA.(U)
MAY 82 D M TITTERINGTON DAAG29-80-C-0041
UNCLASSIFIED MRC-TSR-2376 NL

1 OF 1
AD A
116 190

END
DATE FILMED
07-82
DTIC




MICROCOPY RESOLUTION TEST CHART

MRC Technical Summary Report #2376

# RECURSIVE PARAMETER ESTIMATION USING INCOMPLETE DATA

D. M. Titterington

**Mathematics Research Center**
**University of Wisconsin—Madison**
**610 Walnut Street**
**Madison, Wisconsin 53706**

May 1982

(Received February 26, 1982)



**Approved for public release**
**Distribution unlimited**

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

# RECURSIVE PARAMETER ESTIMATION USING INCOMPLETE DATA

D. M. Titterington*

Technical Summary Report #2376

May 1982

## ABSTRACT

Stochastic approximation procedures are considered for the estimation of parameters using incomplete data. One procedure is stated and illustrated which often leads to asymptotically efficient estimators. Others are developed which, although possibly not optimal in the above sense, will be very much easier to apply. This will be particularly advantageous when quick recursive estimates are required. Examples are given and a link is made between one of the sub-optimal methods and the EM algorithm.

AMS (MOS) Subject Classifications: 62A10, 62F10, 62L12, 62L20.

Key Words: Incomplete data, maximum likelihood estimation, recursive estimation, EM algorithm, stochastic approximation.

Work Unit Number 4 - Statistics and Probability

---

* Department of Statistics, University Gardens, Glasgow G12 8QW, Scotland.

Sponsored by the United States Army under Contract No. DAAG 29-80-C-0041.

# SIGNIFICANCE AND EXPLANATION

Many statistical problems involve the estimation of parameters in a model using data which are incomplete. For instance, some values may be missing altogether or they may be "censored" in that their exact values are not known but are known to fall in a specified range.

Almost without fail, estimation using such data is significantly more awkward than if they were complete and, although numerical methods are available, there is scope for faster procedures, even if the resulting estimates may not be quite as "optimal". This paper describes methods which incorporate the data one at a time into the estimation procedure. This leads to recursive estimates which may well be desirable in themselves, if the data do arrive sequentially. The procedures described are of the "stochastic approximation" type, for which extensive theory exists.

Most emphasis is placed on two such recursions, one which is asymptotical optimal and one which, although suboptimal, will be very much simpler from a computational point of view. This latter method can also be neatly linked to one of the main procedures for nonrecursive estimation in incomplete data problems, the EM algorithm.

A few illustrative examples are given.



The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

# RECURSIVE PARAMETER ESTIMATION USING INCOMPLETE DATA

D. M. Titterington*

## 1. INTRODUCTION

Parameter estimation using incomplete data tends to be much more awkward than with a corresponding set of complete data. Maximum likelihood estimation, for instance, usually requires numerical methods, such as the methods of Scoring and Newton Raphson. Dempster et al (1977) give a compendium of incomplete data problems and describe an alternative numerical iterative procedure, the EM algorithm, which has the mixed blessings of being of first order but monotonic and easy to program. If very large data-sets are involved, then numerical procedures can become very expensive. Their application to survey data with nonresponse could be a case in point.

We shall illustrate here some alternative recursive procedures in which the data are run through once, sequentially. Such a procedure will take the form

$$\underline{\theta}^*_{k+1} = G_k(\underline{\theta}^*_k, y_{k+1}), \; k = 0,1, \ldots \tag{1}$$

where $\underline{\theta}$ denotes the parameter(s). $\underline{\theta}^*_k$ denotes the estimate after $k$ observations and $y_{k+1}$ denotes the $(k+1)$st observation. If there are $n$ observations altogether, then the estimate we would quote is $\underline{\theta}^*_n$.

When data do arrive sequentially, as in control engineering contexts, such recursive procedures may be essential to give "quick" up-to-date parameter estimates, particularly if sequential design is to be incorporated; see Chapter 7 of Goodwin and Payne (1977), Titterington (1980) and references therein. In the more usual statistical contexts, we shall have to impose some

* Department of Statistics, University Gardens, Glasgow G12 8QW, Scotland.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

ordering on the data, in conflict, say, with the likelihood principle (Anderson, 1979). We shall show that, asymptotically, the ordering is irrelevant and, in a later paper (Titterington and Jiang, 1982), evidence will be presented that the ordering effect is not very important in moderately sized samples.

Some simple sequential estimation procedures do not suffer from the criticism of order dependence, as is shown by the following illustrations in which there is no incomplete data.

Example 1.1. Independent Bernoulli trials

Suppose $y_1, y_2, \ldots$ are independent and that $P(y_k=1) = \theta = 1-P(y_k=0)$, $k = 1,2, \ldots$ . Then the recursion

$$\hat{\theta}_{k+1} = \hat{\theta}_k + (k+1)^{-1}(y_{k+1} - \hat{\theta}_k), \; k = 0,1, \ldots$$
$$\hat{\theta}_0 = 0 \quad ,$$

generates exactly the MLE's of $\theta$ as the data are incorporated.

Example 1.2. Exponential-family type models

Suppose $y_1, y_2, \ldots$ are independent and that each has p.d.f.

$$\log f(y|\underline{\phi}) = \text{const} + \underline{t}(y)^T\underline{\phi} + a(\underline{\phi}) \quad ,$$

where $\underline{\phi}$ is a vector and $\underline{t}(y)$ the vector of sufficient statistics for $\underline{\phi}$.

Let $\underline{\theta} = \underline{\theta}(\underline{\phi}) = E(\underline{t}(y)|\underline{\phi})$.

Then, given $y_1, \ldots, y_k$, $\underline{\hat{\theta}}_k$, the MLE, satisfies

$$k\underline{\hat{\theta}}_k = \sum_{i=1}^{k} \underline{t}_i \quad ,$$

where $\underline{t}_i = \underline{t}(y_i)$, $i = 1, \ldots, k$.

We may calculate $\{\underline{\hat{\theta}}_k\}$ recursively from

$$\underline{\hat{\theta}}_{k+1} = \underline{\hat{\theta}}_k + (k+1)^{-1}(\underline{t}_{k+1} - \underline{\hat{\theta}}_k), \; k = 0,1,\ldots, \; \underline{\hat{\theta}}_0 = 0 \quad .$$

The link between Example 1.2 and recursive-least-squares is clear; see also Harrison and Stevens (1976).

In these examples the recursions simply give a convenient way of calculating the usual estimates, and are unnecessary when considering the asymptotic theory and general performance of the estimators produced. Our objective is to develop a similar approach to cope with the possibility of incompleteness in the observations.

## 2. SOME RECURSIVE PROCEDURES

Suppose $y_1, y_2, \ldots$ are independent observations, each with underlying probability density function (p.d.f.) $g(y|\underline{\theta})$, where $\underline{\theta} \in \Theta \subset \mathbf{R}^s$, for some $s$. Let $\underline{S}(y,\underline{\theta})$ denote the vector of scores. That is,

$$S_j(y,\underline{\theta}) = \frac{\partial}{\partial\theta_j} \log g(y|\underline{\theta}),\ j = 1,\ldots,s \quad .$$

Let $\underline{D}^2(y,\underline{\theta})$ denote the matrix of second derivatives of $\log g(y|\underline{\theta})$ and let $I(\underline{\theta})$ denote the Fisher information matrix corresponding to one observation. It is assumed that all derivatives and expected values exist and that

$$E_{\underline{\theta}}\ \underline{S}(y,\underline{\theta}) = \int \underline{S}(y,\underline{\theta})\ g(y|\underline{\theta})dy = \underline{0} \quad ;$$

$$I(\underline{\theta}) = E_{\underline{\theta}}\{\underline{S}(y,\underline{\theta})\underline{S}^T(y,\underline{\theta})\} = -E_{\underline{\theta}}\ \underline{D}^2(y,\underline{\theta}) \quad .$$

Consider the recursion

$$\underline{\theta}^*_{k+1} = \underline{\theta}^*_k + \{kI(\underline{\theta}^*_k)\}^{-1}\underline{S}(y_{k+1},\underline{\theta}^*_k),\ k = 0,1,\ldots \tag{2}$$

which is recognizable as a stochastic approximation procedure. Under regularity conditions over and above those alluded to above, as $k \to \infty$,

$$\sqrt{k}\ (\underline{\theta}^*_k - \underline{\theta}_0) \to N(\underline{0},\ I(\underline{\theta}_0)^{-1}) \quad , \tag{3}$$

in distribution, where $\underline{\theta}_0$ denotes the true parameter value. This result appears in Sacks (1958), Fabian (1968), Nevel'son and Has'minskii (1973, Chapter 8) and Fabian (1978).

We now state the conditions required for the most useful version of the result in Fabian (1978).

(C1) Continuity.

(i) $\int \{\underline{S}(y,\underline{\delta}) - \underline{S}(y,\underline{\theta})\}^T\{\underline{S}(y,\underline{\delta}) - \underline{S}(y,\underline{\theta})\}g(y|\underline{\theta})dy \to 0$ as $\underline{\delta} \to \underline{\theta}$ in $\Theta$.

(ii) If, as $k \to \infty$, $\underline{\theta}^*_k \to \underline{\theta}_0$, then

$$[I(\underline{\theta}^*_k)]^{-1} \to [I(\underline{\theta}_0)]^{-1} \quad .$$

(C2) "Definiteness".

$$-(\underline{\delta}-\underline{\theta})^T I(\underline{\delta})^{-1}\ \mathbb{E}_{\underline{\theta}}\ \underline{S}(y,\underline{\delta}) > 0 \quad \text{for} \quad \underline{\delta} \neq \underline{\theta} \quad . \tag{4}$$

(C3) Boundedness

$$\mathbb{E}_{\underline{\theta}} \| I(\underline{\delta})^{-1}\underline{S}(y,\underline{\delta})\|^2 < C\{1 + \|\underline{\delta}-\underline{\theta}\|^2\} \quad , \tag{5}$$

where $\|\underline{u}\|^2 = \underline{u}^T\underline{u}$ and $C$ is independent of $\underline{\delta}$.

One further comment must be made which has particular relevance to some of the examples in Section 3, namely that it is assumed, in the theory, that $\underline{\theta}^*_k \in \Theta$, for all $k$. In practice, (2) may have to be modified to ensure this. For instance, if $\theta$ is a probability (see Example 3.3, for instance) an additional constraint should be added, such as: $\epsilon < \theta^*_k < 1-\epsilon$, for all $k$ and some small positive $\epsilon$.

Given all this, (3) is guaranteed.

If (3) holds for (2) then it also will for

$$\underline{\theta}^*_{k+1} = \underline{\theta}^*_k + \{(k+1)I(\underline{\theta}^*_k)\}^{-1}\underline{S}(y_{k+1},\underline{\theta}^*_k),\ k = 0,1,\ldots \quad . \tag{6}$$

It is easy to check that the recursive calculations of the MLE's in Examples 1.1 and 1.2 are special cases of (6).

As we shall see in some of the Examples in Section 3, complications may arise in applying recursions (2) and (6), in the computation and inversion, in the multiparameter case, of $I(\underline{\theta}^*_k)$. Numerical integration is often necessary and the fact that we are dealing with incomplete data will add to the complications. Suppose, with reference to (2), we write

$$V_k = \{kI(\underline{\theta}^*_k)\}^{-1} \quad .$$

Then the following alternatives to $V_k^{-1}$ suggest themselves.

(i) $kI(\underline{\theta}')$, where $\underline{\theta}'$ is an initial parameter estimate or one that is updated infrequently.

(ii) $\sum_{i=1}^{k} J_i(\underline{\theta}_k^*)$, where $J_i(\cdot)$ denotes the sample information matrix from the ith observation.

(iii) $\sum_{i=1}^{k} I(\underline{\theta}_i^*)$.

(iv) $\sum_{i=1}^{k} J_i(\underline{\theta}_i^*)$.

Suggestion (i) corresponds to a familiar modification to the Method of Scoring for obtaining maximum likelihood estimates. Suggestion (ii) is similar to Newton's method for the same purpose. Suggestions (iii) and (iv) would be very useful in providing recursive calculation of the $\{V_k^{-1}\}$. If (iv) is used, for instance, we obtain

$$V_k^{-1} = V_{k-1}^{-1} + I(\underline{\theta}_k^*) \quad . \tag{7}$$

Recursion (2), with exactly this modification, was used by Walker and Duncan (1967) in the recursive estimation of parameters in a linear logistic model for quantal response. In their problem the observations are not identically distributed, so that

$$V_k^{-1} = \sum_{i=1}^{k} I_i(\underline{\theta}_i^*) \quad .$$

They are particularly fortunate, in that each $I_i(\underline{\theta})$ is of rank one so that, given $V_0$, all other $\{V_k\}$ can be obtained without further matrix inversion: see their equation (5.4).

Theoretical and practical investigation of these modifications would be worthwhile.

We shall concentrate, however, on the following modification of (2), which suggests itself especially for incomplete data problems.

$$\underline{\tilde{\theta}}_{k+1} = \underline{\tilde{\theta}}_k + \{kI_c(\underline{\tilde{\theta}}_k)\}^{-1}\underline{S}(y_{k+1},\underline{\tilde{\theta}}_k), \; k = 0,1,\ldots \tag{8}$$

where $I_c(\underline{\theta})$ denotes the Fisher Information matrix corresponding to a <u>complete</u> observation. For future reference we denote by equation (9) the version of (8) corresponding to (3). Although these recursions will not lead to asymptotic efficiency, conditions (4) and (3) sometimes guarantee $\sqrt{n}$-consistency and asymptotic Normality. We extract the following theorem from Sacks (1958) and Fabian (1968). We state the univariate version, for future application to the first three examples in Section 3.

Theorem 1.

Given conditions corresponding to those above and provided $2I(\theta_0)I_c(\theta_0)^{-1} > 1$,

$$\sqrt{k}\,(\tilde{\theta}_k - \theta_0) \to N(0,\ I_c(\theta_0)^{-2}I(\theta_0)/\{2I(\theta_0)I_c(\theta_0)^{-1} - 1\})$$

in distribution as $k \to \infty$.

As will become clear, it does not always happen that $2I(\theta_0) > I_c(\theta_0)$. Suppose

$$0 < \beta < 2I(\theta_0)/I_c(\theta_0) < 1$$

and we consider the recursion

$$\tilde{\theta}_{k+1} = \tilde{\theta}_k + k^{-\frac{1}{2}(1+\beta)}I_c(\tilde{\theta}_k)^{-1}S(y,\tilde{\theta}_k),\ k = 0,1,\ldots \ . \tag{10}$$

Then, according to Fabian (1968),

$$k^{\beta/2}(\tilde{\theta}_k - \theta_0) \to N(0,\ I_c(\theta_0)^{-2}I(\theta_0)/\{2I(\theta_0)I_c(\theta_0)^{-1} - \beta\})$$

in distribution, as $k \to \infty$.

Thus, provided there is some information in the incomplete data ($I(\theta_0) > 0$), a modified version of (8) leads to a consistent, asymptotically Normal estimator.

Multidimensional versions of these results will be required in a complete study of Example 3.4 but this will not be undertaken in the present paper: see Sacks (1958) and Fabian (1968).

The important practical advantage of recursions (8), (9) and (10) is that $I_c(\underline{\theta})$ will usually be much easier to evaluate and, if a matrix, to invert, than $I(\underline{\theta})$.

In the following section we derive versions of some of these recursions for a few simple examples involving incomplete data. As $y_1, y_2, \ldots$ represents a sequence of incomplete observations, so $x_1, x_2, \ldots$ will denote corresponding "complete" versions. Thus, given $y$, $x$ belongs to a subset $X(y)$ of the overall sample space $X$ and, if $f(x|\underline{\theta})$ denotes the p.d.f. of $x$, then

$$g(y|\underline{\theta}) = \int_{X(y)} f(x|\underline{\theta})dx \quad ;$$

See Dempster et al (1977).

## 3. SOME EXAMPLES

Example 3.1. Trinomial with incompletely classified observations.

Independent observations are obtained from a trinomial, with cell probabilities $\frac{1}{2}\theta$, $\frac{1}{2}\theta$, $1-\theta$ $(0 < \theta < 1)$. However, all that is known is whether or not the observation belongs to cell 1 $(x = 1$ as opposed to $x = 2$ or 3). Let

$$y = 1 \quad \text{if} \quad x = 1$$
$$= 0 \quad \text{if} \quad x = 2 \quad \text{or} \quad 3 \quad .$$

Then

$$\log g\,(y|\theta) = y \log(\tfrac{1}{2}\theta) + (1-y)\log(1 - \tfrac{1}{2}\theta) \quad ,$$

$$S(y,\theta) = y/\theta - (1-y)/(2-\theta)$$

and
$$I(\theta) = \theta^{-1}(2-\theta)^{-1} \quad .$$

Recursion (2) is

$$\theta^*_{k+1} = \theta^*_k + k^{-1}\{(2-\theta^*_k)y_{k+1} - \theta^*_k(1-y_{k+1})\} \quad .$$

It is not hard to show that conditions (4) and (5) of Section 2 are satisfied.

Similarly, $I_c(\theta) = \theta^{-1}(1-\theta)^{-1}$ and recursion (8) is

$$\tilde{\theta}_{k+1} = \tilde{\theta}_k + k^{-1}\,\tilde{\theta}_k(1-\tilde{\theta}_k)\{y_{k+1}/\tilde{\theta}_k - (1-y_{k+1})/(2-\tilde{\theta}_k)\} \quad .$$

However, for all $0 < \theta < 1$, $I(\theta)/I_c(\theta) = (1-\theta)/(2-\theta) < \frac{1}{2}$ so Theorem 1 will not hold and $\{\tilde{\theta}_k\}$ is not $\sqrt{k}$-consistent. In spite of this it is possible to establish strong consistency of $\{\tilde{\theta}_k\}$ by appeal to a theorem of Gladyshev (1965). Also, for any $\theta_0 > 0$, a modified recursion of the form (10) can be used to obtain a consistent, asymptotically Normal estimator.

Example 3.2. Censored exponential.

Suppose there is censoring on the right at $t_0$ and

$$\log f(x|\theta) = -\log\theta - x/\theta \qquad (x > 0,\ \theta > 0) \quad .$$

Thus,

$$y = x \quad \text{if} \quad x < t_0 \quad .$$

Otherwise $y$ is the knowledge that "$x > t_0$", so that

$$\log g(y|\theta) = -\log\theta - y/\theta \quad \text{if} \quad x < t_0 \quad ,$$

$$= -t_0/\theta, \quad \text{otherwise} \quad .$$

It turns out that (2) is

$$\theta^*_{k+1} = \theta^*_k + \{k(1 - \exp(-t_0/\theta^*_k))\}^{-1}(y_{k+1} - \theta^*_k) \qquad (x_{k+1} < t_0)$$

$$= \theta^*_k + \{k(1 - \exp(-t_0/\theta^*_k))\}^{-1}t_0, \quad \text{otherwise} \quad ,$$

and $I(\theta) = \{1 - \exp(-t_0/\theta)\}/\theta^2$.

Condition (4) is satisfied, its left hand side being

$$(\delta-\theta)^2(1 - \exp(-t_0/\theta))/(1 - \exp(-t_0/\delta)) \quad .$$

However, the left hand side of (5) is

$$\left(1 - e^{-t_0/\delta}\right)^{-2}\left\{\left(1 - e^{-t_0/\theta}\right)(2\theta^2-2\theta\delta+\delta^2) + 2(\theta-\delta)t_0 e^{-t_0/\theta}\right\} \quad ,$$

which tends to infinity as $\delta \to 0$. If, however, we restrict $\delta > \epsilon > 0$, condition (5) will hold.

Since $I_c(\theta) = \theta^{-2}$, Theorem 1 holds provided $1 - \exp(-t_0/\theta_0) > \frac{1}{2}$, that is, if $t_0 > \theta_0 \log 2$. Recursion (8) is

$$\tilde{\theta}_{k+1} = \tilde{\theta}_k + k^{-1}(y_{k+1} - \tilde{\theta}_k) \qquad (x_{k+1} < t_0)$$

$$= \tilde{\theta}_k + k^{-1}t_0 \qquad \text{(otherwise)} \quad .$$

Again, however, Gladyshev's theorem shows strong consistency of $\{\tilde{\theta}_k\}$ for any $t_0 > 0$. Recursions like (10) may also be considered.

<u>Example 3.3</u>. <u>Estimation of mixing weights</u>.

We consider the case of a mixture of $d$ known densities $\{f_j(\cdot),\ j = 1,\ldots,d\}$.

$$g(y|\underline{\theta}) = \sum_{j=1}^{d-1} \theta_j f_j(y) + \left(1 - \sum_{j=1}^{d-1} \theta_j\right)f_d(y) \quad ,$$

where the $\theta_1,\ldots,\theta_d$ are all nonzero probabilities. Then

$$S_j(y|\theta) = \{f_j(y) - f_d(y)\}/g(y|\underline{\theta}), \quad j = 1,\ldots,d-1 \quad ,$$

$$\underline{D}^2_{jr}(y|\underline{\theta}) = -\{f_j(y) - f_d(y)\}\{f_r(y) - f_d(y)\}/\{g(y)|\underline{\theta})\}^2 \quad ,$$
$$j = 1,\ldots,d-1, \; r = 1,\ldots,d-1 \quad .$$

and

$$I_{jr}(\underline{\theta}) = \int \{f_j(y) - f_d(y)\}\{f_r(y) - f_d(y)\}g(y|\theta)^{-1}dy \quad ,$$
$$j,r = 1,\ldots,k-1 \quad .$$

Verification of the regularity conditions is subsumed in Kazakos (1977) and Smith and Makov (1978). For the special case of $d = 2$, with $\theta_1 = \theta$, we obtain, for (2), as in Kazakos (1977),

$$\theta^*_{k+1} = \theta^*_k + \{kI(\theta^*_k)\}^{-1}\{f_1(y_{k+1}) - f_2(y_{k+1})\}/g(y_{k+1}|\theta^*_k), \; k = 1,2,\ldots,$$

with

$$I(\theta) = \int (f_1(y) - f_2(y))^2 g(y|\theta)^{-1}dy \quad .$$

We maintain our concentration on the case $d = 2$.

Here the incompleteness is caused by ignorance of the source of an observed $y$; is it component 1 or component 2? We may write

$$x = (y,\underline{z}) \quad ,$$

where $\underline{z}^T = (1,0)$ or $(0,1)$ according to the source. Thus

$$\log f(x|\theta) = \underline{z}^T\underline{u}(\theta) + \underline{z}^T\underline{v}(\theta)$$

where

$$\underline{u}^T(\theta) = (\log\theta, \log(1-\theta))$$

and
$$\underline{v}^T(\theta) = (\log f_1(y), \log f_2(y)) \quad .$$

Thus, $I_c(\theta) = 1/\theta(1-\theta)$ and (8) becomes

$$\tilde{\theta}_{k+1} = \tilde{\theta}_k + k^{-1}\,\tilde{\theta}_k(1-\tilde{\theta}_k)\{f_1(y_{k+1}) - f_2(y_{k+1})\}/g(y_{k+1}|\tilde{\theta}_k) \quad . \qquad (11)$$

Asymptotically, if $I(\theta) > \frac{1}{2}I_c(\theta)$, Theorem 1 holds. Otherwise, strong consistency can still be guaranteed (see Makov and Smith (1977), Smith and Makov (1978)) and recursions like (10) may also be used.

Example 3.4. Mixture of two univariate Normals.

Let

$$g(y|\underline{\theta},\underline{\mu},\underline{\phi}) = \theta_1 p(y|\mu_1,\phi_1) + \theta_2 p(y|\mu_2,\phi_2)$$

$$= \theta_1 p_1(y) + \theta_2 p_2(y) \quad ,$$

where $0 < \theta_1 = 1-\theta_2 < 1$ and

$$p(y|\mu,\phi) = (2\pi\phi)^{-1/2} \exp\{-\frac{1}{2}(x-\mu)^2/\phi\} \quad .$$

Then the component of the score vector are

$$\partial \log g(y)/\partial\theta_1 = \{p_1(y) - p_2(y)\}/g(y) \quad ,$$

$$\partial \log g(y)/\partial\mu_j = (y-\mu_j)w_j(y)/\phi_j, \quad j = 1,2,$$

$$\partial \log g(y)/\partial\phi_j = \{(y-\mu_j)^2 - \phi_j\}w_j(y)/2\phi_j^2, \quad j = 1,2,$$

where $w_j(y) = \theta_j p_j(y)/g(y)$, $j = 1,2$.

Note that, for $j = 1,2$, $w_j(y)$ is the conditional probability that an observation comes from component $j$, given its datum value, $y$.

Here we do not go with the details of the verification of conditions (4) and (5). They will be complicated, as is application of the recursion (2), itself, because the information matrix is a complicated matrix, even for univariate mixtures, let along multivariate ones. As in Example 3.3, numerical integration is necessary; see Behboodian (1972).

To point out this awkwardness in application is the main reason for mentioning this example. It motivates strongly the use of recursions like (8). For this we require $I_c(\theta_1,\underline{\mu},\underline{\phi})$.

Again $x = (y,\underline{z})$ and now

$$\log f(x|\underline{\theta},\underline{\mu},\underline{\phi}) = \underline{z}^T\underline{u}(\underline{\theta}) + \underline{z}^T\underline{v}(\underline{\mu},\underline{\phi}) \quad ,$$

where, for instance,

$$v_1(\underline{\mu},\underline{\phi}) = \log p(y|\mu_1,\phi_1) \quad .$$

If the parameters are ordered as $\theta_1, \mu_1, \mu_2, \phi_1, \phi_2$, then

$$I_c(\theta_1, \underline{\mu}, \underline{\phi}) = \operatorname{diag}\{\theta_1^{-1}(1-\theta_1)^{-1}, \theta_1/\phi_1, (1-\theta_1)/\phi_2, \theta_1/2\phi_1^2, (1-\theta_1)/2\phi_2^2\}$$

and recursion (8) becomes very simple, as follows.

$$\tilde{\theta}_1^{(k+1)} = \tilde{\theta}_1^{(k)} + k^{-1}\left(w_1^{(k)}(y_{k+1}) - \tilde{\theta}_1^{(k)}\right)$$

$$\tilde{\mu}_j^{(k+1)} = \tilde{\mu}_j^{(k)} + \{k\theta_j^{(k)}\}^{-1} w_j^{(k)}(y_{k+1})(y_{k+1} - \mu_j^{(k)})$$

$$\tilde{\phi}_j^{(k+1)} = \tilde{\phi}_j^{(k)} + \{k\theta_j^{(k)}\}^{-1} w_j^{(k)}(y_{k+1})\{(y_{k+1} - \mu_j^{(k)})^2 - \phi_j^{(k)}\}$$

where $w_j^{(k)}(y) = \theta_j^{(k)} p(y|\mu_j^{(k)}, \phi_j^{(k)})/g(y|\underline{\theta}^{(k)}, \underline{\mu}^{(k)}, \underline{\phi}^{(k)})$, $j = 1,2$ .

## 4. A CONNECTION WITH THE EM ALGORITHM

As pointed out by Fabian (1978, Section 5.8), there is a strong relationship between recursion (2) and the Method of Scoring. Recursion (8), on the other hand, is similarly linked to the EM algorithm.

Suppose $x_1,\ldots,x_n$ represent $n$ independent complete observations, corresponding to $y_1,\ldots,y_n$. Define

$$Q(\underline{\theta}|\underline{\theta}') = E_{\underline{\theta}'}\{\sum_{i=1}^{n} \log f(x_i|\underline{\theta})\ |y_1,\ldots,y_n\} \quad .$$

The EM algorithm generates a sequence $\{\underline{\theta}_m\}$ of parameter estimates by repeating the following double step.

<u>E-step</u>: Evaluate $Q(\underline{\theta}|\underline{\theta}_m)$.

<u>M-step</u>: Choose $\underline{\theta} = \underline{\theta}_{m+1}$ to maximum $Q(\underline{\theta}|\underline{\theta}_m)$.

Consider the following recursive version.

At stage $k+1$, with current estimate $\tilde{\underline{\theta}}_k$, define

$$L_{k+1}(\underline{\theta}) = E_{\tilde{\underline{\theta}}_k}\{\log f(x_{k+1}|\underline{\theta})\ |y_{k+1}\} + L_k(\underline{\theta}) \quad . \qquad (12)$$

Choose $\underline{\theta} = \tilde{\underline{\theta}}_{k+1}$ to maximize $L_{k+1}(\underline{\theta})$. Finally, estimate $\underline{\theta}_0$ by $\tilde{\underline{\theta}}_n$.

Both the EM algorithm and its recursive version may be used in Bayesian analysis for the computation of posterior modes. In (12) we can initialize using

$$L_0(\underline{\theta}) = \log p(\underline{\theta}) \quad ,$$

where $p(\cdot)$ is the prior density for $\underline{\theta}$, with mode $\underline{\theta}_0$.

<u>Theorem 2</u>. Approximately, given appropriate regularity, recursion (12) can be written as

$$\tilde{\underline{\theta}}_{k+1} = \tilde{\underline{\theta}}_k + \{(k+1)I_c(\tilde{\underline{\theta}}_k)\}^{-1}\underline{S}(y_{k+1},\tilde{\underline{\theta}}_k) \quad ,$$

which is the recursion we called (9) in Section 2.

Proof: To clarify the steps we omit some subscripts and rewrite (12) as

$$L_{k+1}(\underline{\theta}) = E_{\underline{\theta}'}\{\log f(x|\underline{\theta})|y\} + L_k(\underline{\theta}) \quad ,$$

where $\underline{\theta}'$ maximizes $L_k(\underline{\theta})$.

We derive the recursion while showing, inductively, that approximately for $\underline{\theta}$ near $\underline{\theta}'$,

$$L_k(\underline{\theta}) = L_k(\underline{\theta}') - \frac{1}{2}(\underline{\theta}-\underline{\theta}')^T\{kI_c(\underline{\theta}')\}(\underline{\theta}-\underline{\theta}') \quad .$$

For $x \in X(y)$, define the conditional density

$$k(x|y,\theta) = f(x|\theta)/g(y|\theta) \quad .$$

Then by Taylor expansion, approximately,

$$\log f(x|\underline{\theta}) = \log f(x|\underline{\theta}') + (\underline{\theta}-\underline{\theta}')^T D_{\underline{\theta}'} \log f(x|\underline{\theta}')$$

$$+ \frac{1}{2}(\underline{\theta}-\underline{\theta}')D^2_{\underline{\theta}'} \log f(x|\underline{\theta}') \cdot (\underline{\theta}-\underline{\theta}')$$

$$= \log f(x|\underline{\theta}') + (\underline{\theta}-\underline{\theta}')^T\{\underline{S}(y,\underline{\theta}') + D_{\underline{\theta}'} \log k(x|y,\underline{\theta}')\}$$

$$+ \frac{1}{2}(\underline{\theta}-\underline{\theta}')^T D^2_{\underline{\theta}'} \log f(x|\underline{\theta}') \cdot (\underline{\theta}-\underline{\theta}') \quad .$$

Given appropriate regularity,

$$E_{\underline{\theta}'}\{D_{\underline{\theta}'} \log k(x|y,\underline{\theta}')|y\} = \underline{0} \quad ,$$

so that, approximately,

$$L_{k+1}(\underline{\theta}) = E_{\underline{\theta}'}\{\log f(x|\underline{\theta}')|y\} + L_k(\underline{\theta}') + (\underline{\theta}-\underline{\theta}')^T \underline{S}(y,\underline{\theta}')$$
$$- \frac{1}{2}(\underline{\theta}-\underline{\theta}')^T\{(k+1)I_c(\underline{\theta}')\}(\underline{\theta}-\underline{\theta}') \quad . \tag{13}$$

The maximizing $\underline{\theta}$ is

$$\hat{\underline{\theta}} = \underline{\theta}' + \{(k+1)I_c(\underline{\theta}')\}^{-1}\underline{S}(y,\underline{\theta}') \quad , \tag{14}$$

which is the required recursion.

Also, from (13),

$$L_{k+1}(\theta) = c + (\underline{\theta}-\hat{\underline{\theta}})^T \underline{S}(y,\underline{\theta}') - \frac{1}{2}(\underline{\theta}-\hat{\underline{\theta}})^T\{(k+1)I_c(\underline{\theta}')\}(\underline{\theta}-\hat{\underline{\theta}})$$
$$- (\underline{\theta}-\hat{\underline{\theta}})\{(k+1)I_c(\hat{\underline{\theta}})\}(\hat{\underline{\theta}}-\underline{\theta}') \quad ,$$

where $c$ is independent of $\underline{\theta}$ ,

$$= c - \frac{1}{2}(\underline{\theta}-\hat{\underline{\theta}})^T\{(k+1)I_c(\underline{\theta}')\}(\underline{\theta}-\hat{\underline{\theta}}), \quad \text{from (14)} \quad ,$$

$$\approx c - \frac{1}{2}(\underline{\theta}-\hat{\underline{\theta}})^T\{(k+1)I_c(\hat{\underline{\theta}})\}(\underline{\theta}-\hat{\underline{\theta}}) \quad .$$

**Theorem 3.** In exponential family models in which $\underline{\theta}$ is the expected value of the sufficient statistic, the recursion is exact.

Proof: Suppose $\log f(x|\underline{\theta}) = b(x) + \underline{t}^T\underline{\phi}(\underline{\theta}) + a(\underline{\phi}(\underline{\theta}))$ where $\underline{t} = \underline{t}(x)$ is a vector of sufficient statistics and

$$E_{\underline{\theta}}(t) = \underline{\theta} \quad .$$

Then

$$\underline{D}_{\underline{\theta}} \log f(x|\underline{\theta}) = I_c(\underline{\theta})(\underline{t}-\underline{\theta}) \quad .$$

Suppose $\underline{D}_{\underline{\theta}} L_k(\underline{\theta}) = kI_c(\underline{\theta})(\underline{\theta}'-\underline{\theta})$. This certainly holds for $k = 1$. Then the stationarity condition for $L_{k+1}(\underline{\theta})$ is

$$I_c(\hat{\underline{\theta}})(\underline{t}'-\hat{\underline{\theta}}) + kI_c(\hat{\theta})(\underline{\theta}'-\hat{\underline{\theta}}) = \underline{0} \quad , \tag{15}$$

where $\underline{t}' = E_{\underline{\theta}'}\{\log f(x|\underline{\theta})|y\}$. Thus, if all information matrices are nonsingular,

$$I_c(\underline{\theta}')(\underline{t}'-\hat{\underline{\theta}}) + kI_c(\underline{\theta}')(\underline{\theta}'-\hat{\underline{\theta}}) = \underline{0} \quad ,$$

i.e.

$$\hat{\underline{\theta}} = \underline{\theta}' + \{(k+1)I_c(\underline{\theta}')\}^{-1}I_c(\underline{\theta}')(\underline{t}'-\underline{\theta}')$$

$$= \underline{\theta}' + \{(k+1)I_c(\underline{\theta}')\}^{-1}\underline{S}(y|\underline{\theta}') \quad .$$

In fact, from (15), $(k+1)\hat{\underline{\theta}} = \underline{t}' + k\underline{\theta}'$, so that

$$\underline{D}_{\underline{\theta}} L_{k+1}(\underline{\theta}) = (k+1)I_c(\hat{\underline{\theta}})(\hat{\underline{\theta}}-\underline{\theta}) \quad .$$

These results can be illustrated by applying recursion (12) to the examples. In 3.1, 3.2 and 3.3, we obtain exactly the same formulae as with recursion (9). In Example 3.4 the recursion on $\theta_1$ is the same and the others differ very slightly as follows.

$$\mu_j^{(k+1)} = \mu_j^{(k)} + f_j^{(k)}(y_{k+1})(y_{k+1} - \mu_j^{(k)}) \quad ,$$

$$\phi_j^{(k+1)} = \phi_j^{(k)} + f_j^{(k)}(y_{k+1})\}(y_{k+1} - \mu_j^{(k)})^2 - (1 - f_j^{(k)}(y_{k+1}))\phi_j^{(k)}\} \quad ,$$

$j = 1,2,$ where

$$f_j^{(k)}(y) = \{k\theta_j^{(k)} + w_j^{(k)}(y)\}^{-1}w_j^{(k)}(y) \quad .$$

Note that $f_j^{(k)}(y) \approx \{k\theta_j^{(k)}\}^{-1} w_j^{(k)}(y)$, for large $k$.

Bayesian versions of some of these recursions have appeared before: that for Example 3.3 (c.f. (11)) in Makov and Smith (1977) and Smith and Makov (1978); that for Example 3.4 in Titterington (1976).

For the exponential family models considered in Theorem 3 the recursions have particularly simple forms, reminiscent of Example 1.2. Recursion (2) is

$$\underset{\sim}{\theta}^*_{k+1} = \underset{\sim}{\theta}^*_k + \{kI(\underset{\sim}{\theta}^*_k)\}^{-1} I_c(\underset{\sim}{\theta}^*_k)\{E(\underset{\sim}{t}^*_{k+1}|y_{k+1},\underset{\sim}{\theta}^*_k) - \underset{\sim}{\theta}^*_k\} \quad .$$

Recursion (8) is

$$\underset{\sim}{\theta}^*_{k+1} = \underset{\sim}{\theta}^*_k + k^{-1}\{E(\underset{\sim}{t}^*_{k+1}|y_{k+1},\underset{\sim}{\theta}^*_k) - \underset{\sim}{\theta}^*_k\} \quad .$$

## 5. DISCUSSION

Although, whenever it is relevant, recursion (2) is the ideal choice, it is likely to be complicated to apply in large problems. There, the modifications of recursions (8) and (10) promise to be much easier in practice. Only a few examples have been described and, apart from the mixtures problems, no missing data example has been discussed. This is rectified in Titterington and Jiang (1982), with emphasis on exponential family and, in particular, multivariate Normal distributions. There, also, are provided numerical details about the relative performance of some of the procedures, which is an important aspect of the study. As Makov (1980) points out in the context of Example 3.3 with $d = 2$, recursion (2) may be unsatisfactorily unstable, relative to (8) or (9), particularly in the early stages.

We finish with a final comment about the EM algorithm. Recursion (2) is related to the method of Scoring, which generates a sequence of estimates $\{\hat{\underline{\theta}}_m\}$ according to the recursion

$$\hat{\underline{\theta}}_{m+1} = \hat{\underline{\theta}}_m + \{nI(\hat{\underline{\theta}}_m)\}^{-1} \sum_{i=1}^{n} \underline{S}(y_i, \hat{\underline{\theta}}_m), \ m = 0,1,\ldots$$

where $y_1,\ldots,y_n$ denotes $n$ independent observations.

It is easy to show, using the methods of Theorem 2, that the EM algorithm is given, approximately, by

$$\hat{\underline{\theta}}_{m+1} = \hat{\underline{\theta}}_m + \{nI_c(\hat{\underline{\theta}}_m)\}^{-1} \sum_{i=1}^{n} \underline{S}(y_i, \hat{\underline{\theta}}_m), \ m = 0,1, \ .$$

Again, for the exponential family case of Theorem 3, the iteration is exact, although a simpler version, of course, is

$$\hat{\underline{\theta}}_{m+1} = n^{-1} \sum_{i=1}^{n} \underline{t}_i^{(m)}, \ \text{where } \underline{t}_i^{(m)} = E_{\hat{\underline{\theta}}_m}(\underline{t}_i | y_i), \ i = 1,\ldots,n \ .$$

## REFERENCES


ANDERSON, J. A. (1979). Multivariate logistic compounds. Biometrika, 66, 17-26.

BEHBOODIAN, J. (1972). Information matrix for a mixture of two normal distributions. J. Statist. Comput. Simul., 1, 295-314.

DEMPSTER, A. P., LAIRD, N. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. R. Statist. Soc. B, 39, 1-38.

FABIAN, V. (1968). On asymptotic normality in stochastic approximation. Ann. Math. Statist., 39, 1327-1332.

FABIAN, V. (1978). On asymptotically efficient recursive estimation. Ann. Statist., 6, 854-866.

GLADYSHEV, E. G. (1965). On stochastic approximation. Theor. Prob. Applics., 10, 275-278.

GOODWIN, G. C. and PAYNE, R. L. (1977). Dynamic System Identification: Experiment Design and Data Analysis. New York: Academic Press.

HARRISON, P. J. and STEVENS, C. F. (1976). Bayesian forecasting. J. R. Statist. Soc B, 38, 205-247.

KAZAKOS, D. (1977). Recursive estimation of prior probabilities using a mixture. IEEE Trans. Inform. Theory, IT-23, 203-211.

MAKOV, U. E. (1980). On the choice of gain functions in recursive estimation of prior probabilities. IEEE Trans. Inform. Theory, IT-26, 497-498.

MAKOV, U. E. and SMITH, A. F. M. (1977). A Quasi-Bayes unsupervised learning process for priors. IEEE Trans. Inform. Theory, IT-23, 761-764.

NEVEL'SON, M. B. and HAS'MINSKII, R. Z. (1973). Stochastic Approximation and Recursive Estimation. American Math. Soc., Providence, RI.

SACKS, J. (1958). Asymptotic distribution of stochastic approximation procedures. Ann. Math. Statist., 29, 373-405.

SMITH, A. F. M. and MAKOV, U. E. (1978). A Quasi-Bayes sequential procedure for mixtures. J. R. Statist. Soc. B, 40, 106-112.

TITTERINGTON, D. M. (1976). Updating a diagnostic system using unconfined cases. Appl. Statist., 25, 238-247.

TITTERINGTON, D. M. (1980). Aspects of optimal design in dynamic systems. Technometrics, 22, 287-299.

TITTERINGTON, D. M. and JIANG, J-M. (1982). Recursive estimation procedures for missing-data problems. In preparation.

WALKER, S. H. and DUNCAN, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. Biometrika, 54, 167-179.

DMT/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER #2376 | 2. GOVT ACCESSION NO. AD-A116190 | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle) Recursive Parameter Estimation Using Incomplete Data | | 5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period |
| | | 6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s) D. M. Titterington | | 8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics & Probability |
| 11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office, P.O. Box 12211, Research Triangle Park, North Carolina 27709 | | 12. REPORT DATE May 1982 |
| | | 13. NUMBER OF PAGES 20 |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) | | 15. SECURITY CLASS. (of this report) UNCLASSIFIED |
| | | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Incomplete data, maximum likelihood estimation, recursive estimation, EM algorithm, stochastic approximation.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Stochastic approximation procedures are considered for the estimation of parameters using incomplete data. One procedure is stated and illustrated which often leads to asymptotically efficient estimators. Others are developed which, although possibly not optimal in the above sense, will be very much easier to apply. This will be particularly advantageous when quick recursive estimates are required. Examples are given and a link is made between one of the sub-optimal methods and the EM algorithm.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

FILMED

7-8